Instance-based and Feature-based Classification Enhancement for Short & Sparse Texts

Guodong Long

FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

UNIVERSITY OF TECHNOLOGY, SYDNEY

A thesis submitted for the degree of

Doctor of Philosophy

July, 2014

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student

_____

Acknowledgements

First I would like to thank my supervisors, Prof Chengqi Zhang, Prof Xingquan Zhu and Dr Ling Chen. They brought me into the wonderful world of research. They not only gave me valuable advice in academics, but also helped my transition into a different culture. Prof Chengqi Zhang has been a great mentor and collaborator, energetic and full of ideas. Prof Xingquan Zhu guided me further into data mining. I am always impressed by his vigour and sharp thinking. Dr Ling Chen helped me by asking insightful questions, and giving me thoughtful comments on the thesis. I have enjoyed working with all of them, and have benefited enormously from the interactions with them.

I spent nearly four years at the University of Technology, Sydney. I thank the collaborators, faculty members, staff, fellow students and friends in the Centre for Quantum Computation and Intelligent Systems, who made my graduate life a very memorable experience. In particular, I thank you if you are reading this thesis.

Finally I thank my family. My parents endowed me with the gift of curiosity about the natural world. My sister, brother-in-law and parents-in-law gave me so much encouragement. Last but not least, my dear wife Jing brings to life so much love and happiness, making thesis writing an enjoyable endeavour.

Abstract

Short, sparse texts are becoming increasingly prevalent as a result of the growing popularity of social networking web sites, such as micro-blogs, Twitter and Flickr, and sites offering online product reviews. These short & sparse texts usually consist of a dozen or more words, or a few sentences, which we represent as a sparse document-term matrix. Compared to normal texts, short & sparse texts have three specific characteristics: (1) insufficient word co-occurrence to measure similarity, (2) low quality data resulting from spelling error, acronyms and slang, and (3) data sparseness. Normal classification methods therefore fail to achieve the desired level of accuracy for classifying short & sparse text.

In this thesis, we present a series of novel approaches to enhance the performance of short & sparse text classification. Most texts can be represented as a two-dimensional matrix and we use the terms - "instance" and "feature" to denote the "row" and "column" concept respectively in the matrix. Corresponding to the matrix's two dimensions, we design an instance- and feature-based framework to expand the rows/columns in the matrix.

- for the instance-based framework, we extract an auxiliary dataset from an external online source (i.e. Wikipedia) with predefined class information, and integrate the target and auxiliary datasets with an instance-based transfer learning tool to enhance the classification performance of the target short text

domain. Moreover, we propose a sampling framework to handle the challenge of low quality data in auxiliary dataset;

- for the feature-based framework, we infer two kinds of feature sets with the given short texts, and then combine them with multi-view learning tool to enhance the classification performance. To handle the view disagreement challenge, we integrate a Bagging framework with Multi-view learning.

The aim of the proposed algorithms is to improve classification performance (i.e. accuracy). To evaluate the proposed algorithms, we test them using a variety of benchmark datasets and real world datasets, such as sentiment texts in Twitter, preprocessed 20 Newsgroup data, review texts for seminars, and search snippets. Moreover, we compare the algorithm with other benchmark algorithms on all datasets. The results of our experiments demonstrate that the accuracy of our proposed algorithms is superior to that of other similar algorithms.

Contents

List of Figures

List of Tables